Exercise 5

Juuso Korhonen - 652377
ELEC-E8125 - Reinforcement Learning

October 19, 2022

1 Task 1

(a) training performance plot of basic REINFORCE without baseline is in figure 1.

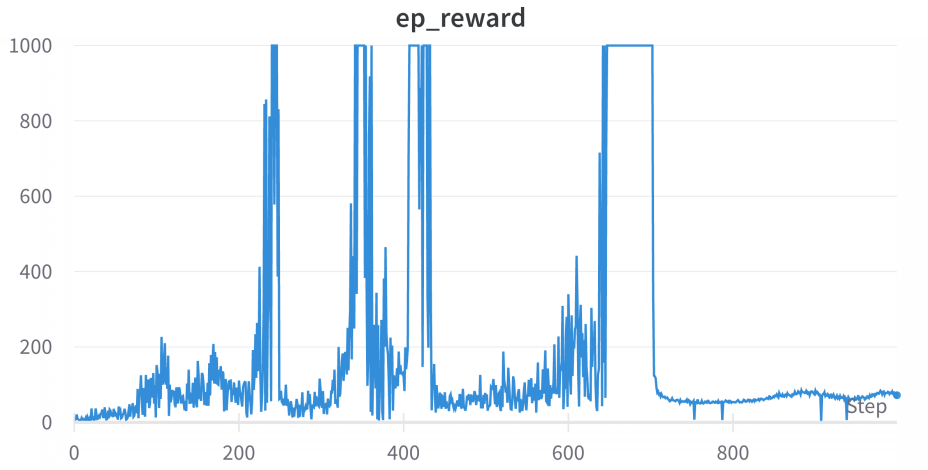

Figure 1: (a) Training performance plot of basic REINFORCE without baseline.

(b) training performance plot of basic REINFORCE with baseline = 20 is in figure 2.
(c) training performance plot of REINFORCE with discounted rewards normalized to zero mean and unit variance is in figure 3.

2 Question 1.1

Q: How would you choose a good value for the baseline? Why is the training more stable when using a baseline?
A: It should be chosen as the mean of the discounted rewards as this shows good performance

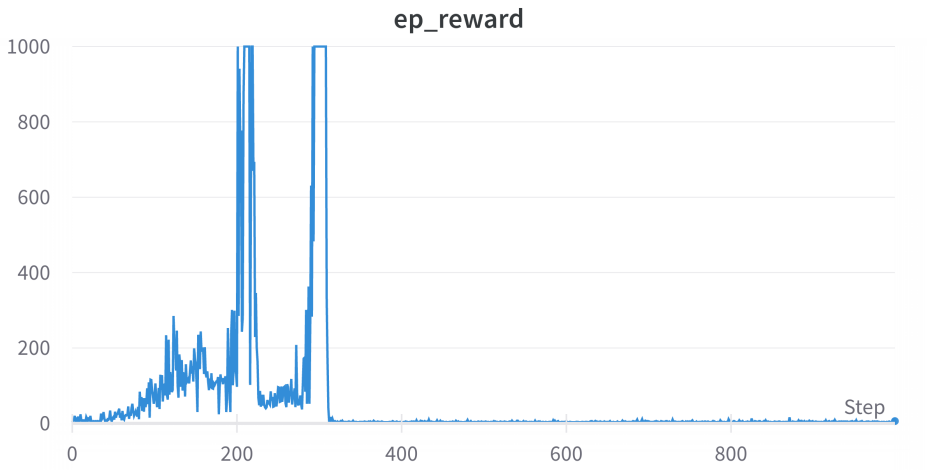

Figure 2: (b) Training performance plot of basic REINFORCE with baseline = 20.

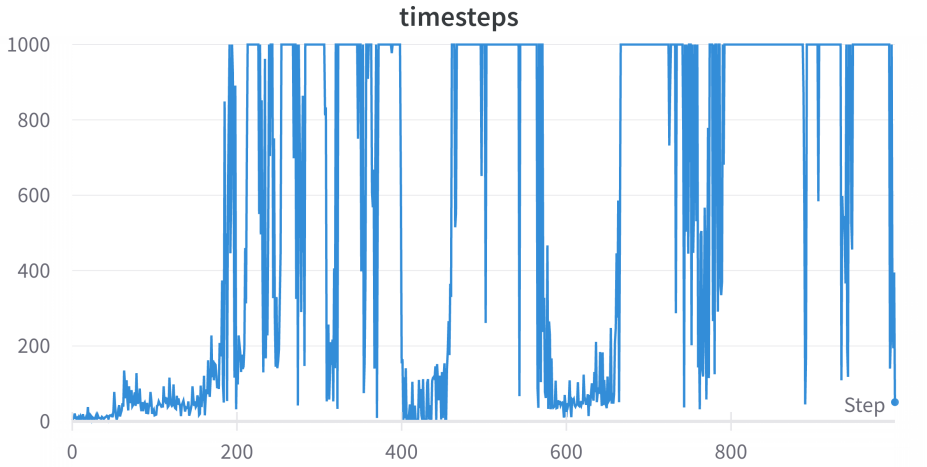

Figure 3: (c) Training performance plot of REINFORCE with discounted rewards normalized to zero mean and unit variance.

in (c). The training is more stable using baseline, because the variance of the gradient ends up being lower.

3 Task 2

Training performance plot of REINFORCE with discounted rewards normalized to zero mean and unit variance and learnable variance parameter is in figure 4.

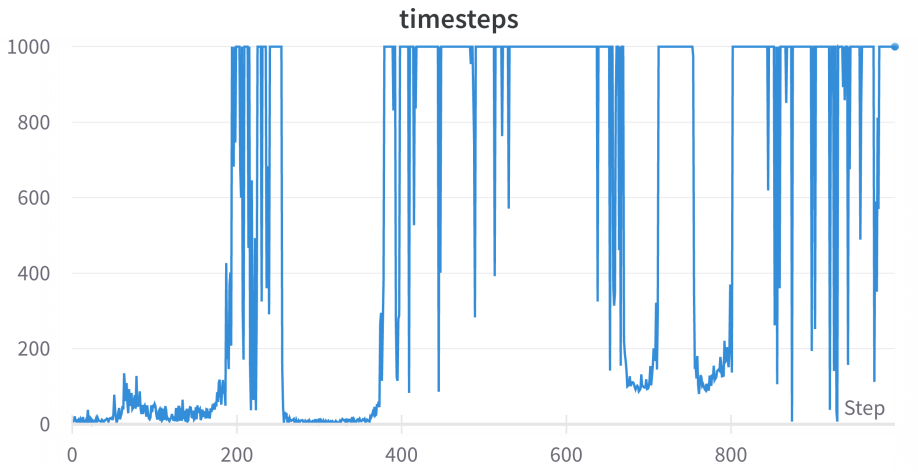

Figure 4: (c) Training performance plot of REINFORCE with discounted rewards normalized to zero mean and unit variance and learnable parameter.

4 Question 2.1

Benefits of constant variance is that it can encourage exploration if set suitably. Downsides are that it has to be configured by hand (error-prone) and learning can take a really long time if the sampling of the action distribution produces suboptimal actions due to setting too high variance.

Benefits of learnable variance is that the agent can better exploit the knowledge it has learned and learning an adequate solution can be faster this way. Downsides are that the agent can learn to be confident of its policy too early and stop exploring too soon (thus not learning perhaps a better policy).

5 Question 2.2

Setting too low initial variance can cause the agent to exploration too low.

6 Question 3

Q: Why the method implemented in this exercise could not be directly used with experience replay? Which steps of the algorithm would be problematic to perform with experience replay? How the problematic steps could be resolved? Explain your answer.

A: In REINFORCE we assume that the sampling of the actions is from our optimized policy. This however can be solved using importance sampling.

7 Question 4.1

Q: What could go wrong when a model with an unbounded continuous action space and a reward function like the one used here (+1 for survival) were to be used with a physical system?

A: Unbounded action space could break the physical system. Also the agent could ask the system to perform impossible actions (instantaneous and out of bounds).

8 Question 4.2

Q: How could the problems appearing in Question 4.1 be mitigated without putting a hard limit on the actions?
A: We could modify the loss to include some kind of cost for the action (for example magnitude).

9 Discrete Action Spaces

Q: Can policy gradient methods be used with discrete action spaces? Why/why not? Which steps of the algorithm would be problematic to perform, if any?

Formulating a probability distribution for discrete actions could be problematic, but could be done for example with multivariate Bernoulli. This is however, computationally intensive.